Available on CMS information server

CMS CR 2007/056

CMS Conference Report

October 3, 2007

The CMS DAQ and Run Control System

Alexander Oh

CMS Collaboration

Abstract

This paper discusses the design and implementation of the Data Acquisition and Run Control systems of the CMS experiment and their current status of installation and commissioning.

Presented at *The 2007 Europhysics Conference on High Energy Physics,* Manchester, UK, July 21, 2007

Submitted to *IoP Journals*

The CMS DAQ and Run Control System
Journal of Physics: Conference Series

Alexander Oh
CERN, Geneva, Switzerland
on behalf of the CMS DAQ Group

Abstract This paper discusses the design and implementation of the Data Acquisition and Run Control systems of the CMS experiment and their current status of installation and commissioning.

1. Introduction
The Data Acquisition (DAQ) [1] has the task to transport the data from about 650 front ends at the detector side, to the "filter units" for processing of complete events. The central DAQ runs online software on about 3000 PC used for intelligent buffering and processing of event data.

The purpose of the Run Control system is to configure and control the central DAQ system and to orchestrate the sub-detector and the first level trigger systems to allow efficient data taking. The Run Control system is based on a common framework and definitions of finite state machines.

Currently, the DAQ and Run Control are in the commissioning phase. With the final hardware and the software being installed and tested. In the following the architecture and the technologies used in the DAQ and Run Control components of CMS will be discussed.

2. The DAQ
The principle components of the DAQ system of CMS are shown in Figure 1. The detector front ends are read out through a builder network with a bisectional bandwidth of 100 GB/s. Complete events are fed to the event filter systems at a rate of maximal 100 kHz. The large rate to the filter systems stems from the design choice of CMS to build the full event already after the first level trigger instead of building partial events as in traditional multi level trigger systems. This requires the read-out, assembly and forwarding of the full event data at the nominal level one trigger rate.

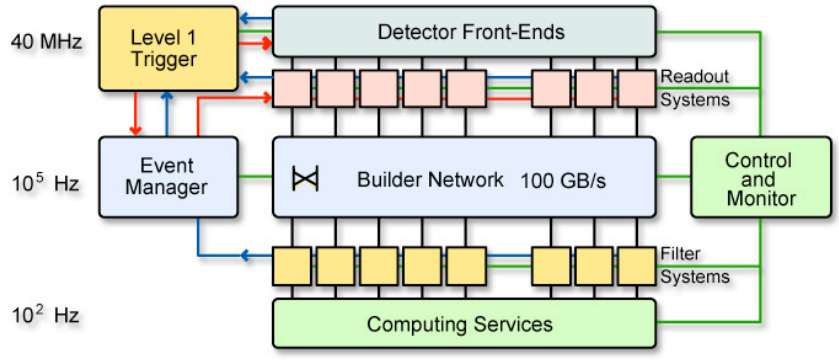

Figure 1 The principal components of the DAQ system of CMS.

2.1. Super-fragment Builder

The event building in the builder network is accomplished in two stages. The first stage, the super-fragment builder, assembles typically 8 fragments into one super-fragment with an average size of 16 kB. The super-fragment builder is based on Myrinet [2] technology. The optical cables of the super fragment builder measure about each 200 m in length to bridge the front-end read-out in the underground cavern to the counting room on the surface. The switches in the super-fragment builder consist of six chassis housing up to 20 line cards each with 16 ports. The cabling layout of the switches has been chosen such that super fragments can be routed to any output destination, allowing greatest flexibility in the configuration of the DAQ.

The performance of the super-fragment builder has been studied using a prototype builder and generating varying fragment size distributions. A rate of 300 MB/s for a log-normal distribution of fragment sizes with an RMS and mean of 2 kB was measured, which exceeds the requirement of 200 MB/s at a nominal fragment size of 2 kB.

2.2. Event Builder

The second stage of the event builder assembles super-fragments into complete events. It consists of PCs housing both a Myrinet card and a four-port Gigabit Ethernet (GBE) card connected to two Force 10 E12000 switches. The Myrinet NIC receives super-fragments from the first stage event builder and the full event is transmitted using TCP/IP over GBE to the Filter Nodes for processing. A test with the final hardware has shown that the requirement of 200 MB/s throughput per input node is easily attained for an assumed average super fragment size of 16 kB.

Because the first stage super-fragment building has an eight-fold symmetry, up to eight second stage event builders can be easily added to the system. The two-stage architecture of the event builder allows staging the deployment of the DAQ while providing the full functionality of the event-builder. This architecture is also inherently stable against failure in the second stage since a catastrophic failure within any of the parallel readout slices, while leading to a performance decrease of $1/8^{th}$ of the full system, will not impact the functionality of the system.

2.3. Deployment

For the first Physics Run foreseen to take place in mid 2008 an Event Builder and Filter Farm with 50% nominal capacity is foreseen. The capacity Event Builder will be gradually expanded reaching 100% for the high luminosity operation of the LHC.

As of today the first stage of the event builder is completed in terms of hardware purchasing and installation. For the second stage hardware for the CMS start up event building is purchased and installed, however purchasing the full processing capability for the start-up system will be done as late as possible to take full advantage of Moore's law in computing capability.

3. The Run Control

The Run Control System configures and controls the online applications of the DAQ components and the Detector Control Systems. It is an interactive system and provides diagnostic information [3]. There are about 10000 applications to manage, running on about 3000 PCs. The Run Control and online components are embedded in a common distributed processing environment (Xdaq [4]). The Run Control structure is organized into eleven different sub-systems, with each sub-system corresponding to a sub-detector, e.g. the Hadron Calorimeter, central DAQ or global trigger. A framework (Run Control and Monitoring System, RCMS) provides a uniform API to common tasks like storage and retrieval from the process configuration DB, state-machine models for process control, and access to the monitoring system.

3.1. Architecture and Technology

A tree of finite state machines implemented with RCMS controls the data taking operation of the experiment. The so-called "Function Manager" (FM) consists of a finite state machine and a set of

services and is the basic element in the control tree. The state machine model has been standardized for the first level of FM's in the control. The advantages of common interfaces and software infrastructures across the sub-detector Run Control and DAQ components facilitate the integration into the central Run Control system.

The services are implemented in Java 1.5.0 as Web Applications. The interfaces are specified with the Web Service Description Language (WSDL) using the Axis implementation of Web Services (WS). Tested Web Service clients include Java, LabView and Perl. The application is running in the publicly available official reference implementation of the Java Servlet technology Tomcat 5 by the Apache Software Foundation. For persistency both Oracle 10g and MySQL 5 are supported by RCMS.

Currently the services and tools provided by RCMS comprise a security service for authentication and user account management, a resource service for storing and delivering configuration information of online processes, a Function Manager Framework providing an API to implement finite state machines, access to remote processes via facades, error handlers, a log message application to collect, store and distribute messages, and Job Control to start, stop and monitor processes in a distributed environment. Furthermore an application has been developed to generate configurations for the DAQ applications.

3.2. Deployment

The PCs to run RCMS have been installed and are being used in commissioning runs with a set of sub-detectors. Ten PCs running Scientific Linux are sufficient to control the experiment. One common database (Oracle 10g) is shared by all online processes and RCMS installations. The RCMS installations of each sub-detector use separate accounts for the Resource Service. Configuration management across sub-systems is achieved using global configuration keys.

The Run Control system has been used successfully to control the Trigger, DAQ and a set of sub-detectors in a global running mode. The RCMS and DAQ commissioning plan foresees regular global runs with a gradually increasing number of sub-systems and number of nodes participating. These global runs provide valuable input to improve and refine the Run Control system of CMS to achieve the highest possible data taking efficiency.

4. Summary

The DAQ and Run Control system has been discussed. The key aspects of the DAQ are the full read-out of the complete event at the level one trigger rate and the flexible two stage design to optimize deployment and usage. A substantial part of the hardware has been purchased and is installed. The Run Control system has been designed as a tree of standardized finite state machines and the Run Control services are implemented using Web service technologies.

Sub-detectors are currently being integrated into the central systems and global runs using both cosmic muon and random trigger sources are being used to commission the CMS detector and DAQ systems in order to be ready for the first pp collisions of LHC, currently scheduled for the end of 2008.

[1] The CMS Collaboration 2002, CMS, The TriDAS Project, Technical Design Report, Volume 2: *Data Acquisition and High-Level Trigger*, CERN LHCC 2002-36

[2] http://www.myri.com/

[3] E. Frizziero, M. Gulmini, F. Lelli, G. Maron, A. Oh, S. Orlando, A. Petrucci, S. Squizzato, and S. Traldi, *Instrument Element: A New Grid component that Enables the Control of Remote Instrumentation* in Proceedings of the Sixth IEEE international Symposium on Cluster Computing and the Grid (Ccgrid'06) - Volume 00/ (May 16 - 19, 2006). CCGRID. IEEE Computer Society, Washington, DC, 52. , 2006.

[4] J. Gutleber and L. Orsini 2002 *Software architecture for processing clusters based on i2o* (Cluster Computing 5(1)) pp 55-64